Automatically Assessing the Post Quality in Online Discussions on Software

Markus Weimer and **Iryna Gurevych** and **Max Mühlhäuser**
Ubiquitous Knowledge Processing Group, Division of Telecooperation
Darmstadt University of Technology, Germany
http://www.ukp.informatik.tu-darmstadt.de
[mweimer,gurevych,max]@tk.informatik.tu-darmstadt.de

Abstract

Assessing the quality of user generated content is an important problem for many web forums. While quality is currently assessed manually, we propose an algorithm to assess the quality of forum posts automatically and test it on data provided by Nabble.com. We use state-of-the-art classification techniques and experiment with five feature classes: Surface, Lexical, Syntactic, Forum specific and Similarity features. We achieve an accuracy of 89% on the task of automatically assessing post quality in the software domain using forum specific features. Without forum specific features, we achieve an accuracy of 82%.

1 Introduction

Web 2.0 leads to the proliferation of user generated content, such as blogs, wikis and forums. Key properties of user generated content are: low publication threshold and a lack of editorial control. Therefore, the quality of this content may vary. The end user has problems to navigate through large repositories of information and find information of high quality quickly. In order to address this problem, many forum hosting companies like Google Groups[1] and Nabble[2] introduce rating mechanisms, where users can rate the information manually on a scale from 1 (low quality) to 5 (high quality). The ratings have been shown to be consistent with the user community by Lampe and Resnick (2004). However, the percentage of manually rated posts is very low (0.1% in Nabble).

[1] http://groups.google.com
[2] http://www.nabble.com

Departing from this, the main idea explored in the present paper is to investigate the feasibility of automatically assessing the perceived quality of user generated content. We test this idea for online forum discussions in the domain of software. The *perceived quality* is not an objective measure. Rather, it models how the community at large perceives post quality. We choose a machine learning approach to automatically assess it.

Our main contributions are: (1) An algorithm for automatic quality assessment of forum posts that learns from human ratings. We evaluate the system on online discussions in the software domain. (2) An analysis of the usefulness of different classes of features for the prediction of post quality.

2 Related work

To the best of our knowledge, this is the first work which attempts to assess the quality of forum posts automatically. However, on the one hand work has been done on automatic assessment of other types of user generated content, such as essays and product reviews. On the other hand, student online discussions have been analyzed.

Automatic text quality assessment has been studied in the area of automatic essay scoring (Valenti et al., 2003; Chodorow and Burstein, 2004; Attali and Burstein, 2006). While there exist guidelines for writing and assessing essays, this is not the case for forum posts, as different users cast their rating with possibly different quality criteria in mind. The same argument applies to the automatic assessment of product review usefulness (Kim et al., 2006c):

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 125–128,
Prague, June 2007.

Stars	Label on the website	Number
⋆	Poor Post	1251
⋆⋆	Below Average Post	44
⋆⋆⋆	Average Post	69
⋆⋆⋆⋆	Above Average Post	183
⋆⋆⋆⋆⋆	Excellent Post	421

Table 1: Categories and their usage frequency.

Readers of a review are asked "Was this review helpful to you?" with the answer choices Yes/No. This is very well defined compared to forum posts, which are typically rated on a five star scale that does not advertise a specific semantics.

Forums have been in the focus of another track of research. Kim et al. (2006b) found that the relation between a student's posting behavior and the grade obtained by that student can be assessed automatically. The main features used are the number of posts, the average post length and the average number of replies to posts of the student. Feng et al. (2006) and Kim et al. (2006a) describe a system to find the most authoritative answer in a forum thread. The latter add speech act analysis as a feature for this classification. Another feature is the author's trustworthiness, which could be computed based on the automatic quality classification scheme proposed in the present paper. Finding the most authoritative post could also be defined as a special case of the quality assessment. However, it is definitely different from the task studied in the present paper. We assess the perceived quality of a given post, based solely on its intrinsic features. Any discussion thread may contain an indefinite number of good posts, rather than a single authoritative one.

3 Experiments

We seek to develop a system that adapts to the quality standards existing in a certain user community by learning the relation between a set of features and the perceived quality of posts. We experimented with features from five classes described in table 2: *Surface, Lexical, Syntactic, Forum specific and Similarity features.*

We use forum discussions from the *Software* category of Nabble.com.[5] The data consists of 1968 rated posts in 1788 threads from 497 forums. Posts can be rated by multiple users, but that happens rarely. 1927 posts were rated by one, 40 by two and 1 post by three users. Table 1 shows the distribution of average ratings on a five star scale. From this statistics, it becomes evident that users at Nabble prefer extreme ratings. Therefore, we decided to treat the posts as being binary rated.: Posts with less than three stars are rated "bad". Posts with more than three stars are "good".

We removed 61 posts where all ratings are exactly three stars. We removed additional 14 posts because they had contradictory ratings on the binary scale. Those posts were mostly spam, which was voted high for commercial interests and voted down for being spam. Additionally, we removed 30 posts that did not contain any text but only attachments like pictures. Finally, we removed 331 non English posts using a simple heuristics: Posts that contained a certain percentage of words above a pre-defined threshold, which are non-English according to a dictionary, were considered to be non-English.

This way, we obtained 1532 binary classified posts: 947 good posts and 585 bad posts. For each post, we compiled a feature vector, and feature values were normalized to the range $[0.0, \ldots, 1.0]$.

We use support vector machines as a state-of-the-art-algorithm for binary classification. For all experiments, we used a C-SVM with a gaussian RBF kernel as implemented by LibSVM in the YALE toolkit (Chang and Lin, 2001; Mierswa et al., 2006). Parameters were set to $C = 10$ and $\gamma = 0.1$. We performed stratified ten-fold cross validation[6] to estimate the performance of our algorithm. We repeated several experiments according to the leave-one-out evaluation scheme and found comparable results to the ones reported in this paper.

4 Results and Analysis

We compared our algorithm to a majority class classifier as a baseline, which achieves an accuracy of 62%. As it is evident from table 3, most system configurations outperform the baseline system. The best performing single feature category are the Forum specific features. As we seek to build an adaptable system, analyzing the performance without these features is worthwhile: Using all other features, we

[5] `http://www.nabble.com/Software-f94.html`

[6] See (Witten and Frank, 2005), chapter 5.3 for an in-depth description.

Feature category	Feature name	Description
Surface Features	Length	The number of tokens in a post.
	Question Frequency	The percentage of sentences ending with “?”.
	Exclamation Frequency	The percentage of sentences ending with “!”.
	Capital Word Frequency	The percentage of words in CAPITAL, which is often associated with shouting.
Lexical Features Information about the wording of the posts	Spelling Error Frequency	The percentage of words that are not spelled correctly.[3]
	Swear Word Frequency	The percentage of words that are on a list of swear words we compiled from resources like WordNet and Wikipedia[4], which contains more than eighty words like “asshole”, but also common transcriptions like “f*ckin”.
Syntactic Features		The percentage of part-of-speech tags as defined in the PENN Treebank tag set (Marcus et al., 1994). We used TreeTagger (Schmid, 1995) based on the english parameter files supplied with it.
Forum specific features Properties of a post that are only present in forum postings	IsHTML	Whether or not a post contains HTML. In our data, this is encoded explicitly, but it can also be determined by regular expressions matching HTML tags.
	IsMail	Whether or not a post has been copied from a mailing list. This is encoded explicitly in our data.
	Quote Fraction	The fraction of characters that are inside quotes of other posts. These quotes are marked explicitly in our data.
	URL and Path Count	The number of URLs and filesystem paths. Post quality in the software domain may be influenced by the amount of tangible information, which is partly captured by these features.
Similarity features		Forums are focussed on a topic. The relatedness of a post to the topic of the forum may influence post quality. We capture this relatedness by the cosine between the posts unigram vector and the unigram vector of the forum.

Table 2: Features used for the automatic quality assessment of posts.

achieve an only slightly worse classification accuracy. Thus, the combination of all other features captures the quality of a post fairly well.

SUF	LEX	SYN	FOR	SIM	Avg. accuracy
		Baseline			*61.82%*
√	√	√	√	√	89.10%
√	–	–	–	–	61.82%
–	√	–	–	–	71.82%
–	–	√	–	–	82.64%
–	–	–	√	–	85.05%
–	–	–	–	√	62.01%
–	√	√	√	√	89.10%
√	–	√	√	√	89.36%
√	√	–	√	√	85.03%
√	√	√	–	√	82.90%
√	√	√	√	–	88.97%
–	√	√	√	–	88.56%
√	–	–	√	–	85.12%
–	–	√	√	–	88.74%

Table 3: Accuracy with different feature sets. SUF: Surface, LEX: Lexical, SYN: Syntax, FOR: Forum specific, SIM: similarity. The *baseline* results from a majority class classifier.

We performed additional experiments to identify the most important features from the Forum specific ones. Table 4 shows that IsMail and Quote Fraction are the dominant features. This is noteworthy, as those features are not based on the domain of discussion. Thus, we believe that these features will perform well in future experiments on other data.

ISM	ISH	QFR	URL	PAC	Avg. accuracy
√	√	√	√	√	*85.05%*
√	–	–	–	–	73.30%
–	√	–	–	–	61.82%
–	–	√	–	–	73.76%
–	–	–	√	–	61.29%
–	–	–	–	√	61.82%
–	√	√	√	√	74.41%
√	–	√	√	√	*85.05%*
√	√	–	√	√	73.30%
√	√	√	–	√	*85.05%*
√	√	√	√	–	*85.05%*
√	–	√	–	–	84.99%
√	√	√	–	–	*85.05%*

Table 4: Accuracy with different forum specific features. ISM: IsMail, ISH: IsHTML, QFR: QuoteFraction, URL: URLCount, PAC: PathCount.

Error Analysis Table 5 shows the confusion matrix of the system using all features. Many posts that were misclassified as good ones show no apparent reason to be classified as bad posts to us. The understanding of their rating seems to require deep knowledge about the specific subject of discussion. The few remaining posts are either spam or rated negatively to signalize dissent with the opinion expressed in the post. Posts that were misclassified as bad ones often contain program code, digital signatures or other non-textual parts in the body. We plan to address these issues with better preprocessing in

	true good	true bad	sum
pred. good	490	72	562
pred. bad	95	875	970
sum	585	947	1532

Table 5: Confusion matrix for the system using all features.

the future. However, the relatively high accuracy already achieved shows that these issues are rare.

5 Conclusion and Future Work

Assessing post quality is an important problem for many forums on the web. Currently, most forums need their users to rate the posts manually, which is error prone, labour intensive and last but not least may lead to the problem of premature negative consent (Lampe and Resnick, 2004).

We proposed an algorithm that has shown to be able to assess the quality of forum posts. The algorithm applies state-of-the-art classification techniques using features such as *Surface, Lexical, Syntactic, Forum specific and Similarity features* to do so. Our best performing system configuration achieves an accuracy of 89.1%, which is significantly higher than the baseline of 61.82%. Our experiments show that forum specific features perform best. However, slightly worse but still satisfactory performance can be obtained even without those.

So far, we have not made use of the structural information in forum threads yet. We plan to perform experiments investigating speech act recognition in forums to improve the automatic quality assessment. We also plan to apply our system to further domains of forum discussion, such as the discussions among active Wikipedia users.

We believe that the proposed algorithm will support important applications beyond content filtering like automatic summarization systems and forum specific search.

Acknowledgments

This work was supported by the German Research Foundation as part of the Research Training Group "Feedback-Based Quality Management in eLearning" under the grant 1223. We are thankful to Nabble for providing their data.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning, and Assessment*, 4(3), February.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-raters performance on toefl essays. Technical report, ETS.

Donghui Feng, Erin Shaw, Jihie Kim, and Eduard Hovy. 2006. Learning to detect conversation focus of threaded discussions. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NNACL)*.

Jihie Kim, Grace Chern, Donghui Feng, Erin Shaw, and Eduard Hovya. 2006a. Mining and assessing discussions on the web through speech act analysis. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference*.

Jihie Kim, Erin Shaw, Donghui Feng, Carole Beal, and Eduard Hovy. 2006b. Modeling and assessing student activities in on-line discussions. In *Proceedings of the Workshop on Educational Data Mining at the conference of the American Association of Artificial Intelligence (AAAI-06)*, Boston, MA.

Soo-Min Kim, Patrick Pantel, Tim Chklovski, and Marco Pennacchiotti. 2006c. Automatically assessing review helpfulness. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 423 – 430, Sydney, Australia, July.

Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: Distributed moderation in a large online conversation space. In *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems, Vienna Austria*, pages 543–550.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. YALE: Rapid prototyping for complex data mining tasks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA. ACM Press.

Helmut Schmid. 1995. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, Manchester, UK.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2:319–329.

Ian H. Witten and Eibe Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2 edition.